Introduction to Machine Learning

Lecture 4: Linear Algebra

Prof. Subrahmanya Swamy Peruru
Department of Electrical Engineering
IIT Kanpur

Recap

Law of Total Expectation

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]]$$

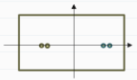
Interpretation: overall average = *weighted average of group averages*

Marks of 5 students: 10,20,30,40,50.

Section split:
Section 1 has 2 students: marks 10,20
Section 2 has 3 students: marks 30,40,50

Law of Total Variance

$$\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(\mathbb{E}[X \mid Y])$$



Interpretation: total spread = (average spread within sections) + (spread of section averages).

Random vector and mean vector

Random vector. Let

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \quad \text{(each } X_i \text{ is a random variable).}$$

Mean / expectation vector.

$$\mathbb{E}[\mathbf{X}] = \begin{bmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_n] \end{bmatrix}$$

Linearity (component-wise). For fixed matrices/vectors and random vectors $\mathbf{X}, \mathbf{Z}$:

$$\boxed{\mathbb{E}[\mathbf{X} + \mathbf{Z}] = \mathbb{E}[\mathbf{X}] + \mathbb{E}[\mathbf{Z}]}$$

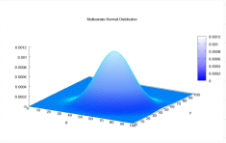
$$\boxed{\mathbb{E}[A\mathbf{X} + b] = A\,\mathbb{E}[\mathbf{X}] + b}$$

Gaussian random vector: pdf

A random vector $\mathbf{X} \in \mathbb{R}^n$ is **Gaussian** (multivariate normal) with mean $\boldsymbol{\mu}$ and covariance Σ if

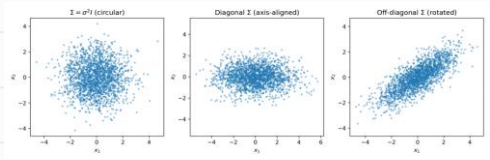
$$\boxed{f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}$$

$$\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}]$$
$$\Sigma = \mathrm{Cov}(\mathbf{X})$$



Key fact: Gaussian is completely characterized by $(\boldsymbol{\mu}, \Sigma)$.

Covariance controls contour shape



$$\Sigma = \begin{bmatrix} 1.5 & 0 \\ 0 & 1.5 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 3 & 0 \\ 0 & 0.8 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 2 & 1.2 \\ 1.2 & 1.2 \end{bmatrix}$$

Cross entropy: Numerical example

Numerical example.
Let

$$p = [0.8,\ 0.2], \qquad q_1 = [0.75,\ 0.25], \qquad q_2 = [0.4,\ 0.6].$$

Cross entropy with q_1:

$$\begin{aligned} H(p, q_1) &= -(0.8\log(0.75) + 0.2\log(0.25)) \\ &\approx -(0.8(-0.288) + 0.2(-1.386)) \\ &\approx 0.51 \end{aligned}$$

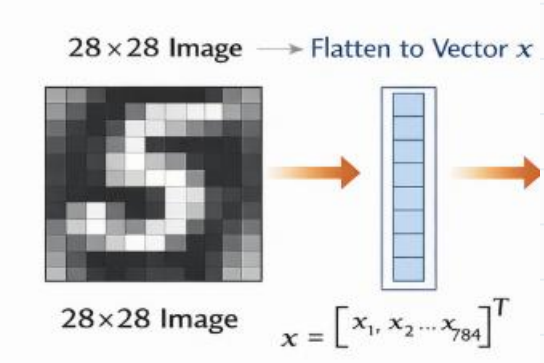
Cross entropy with q_2:

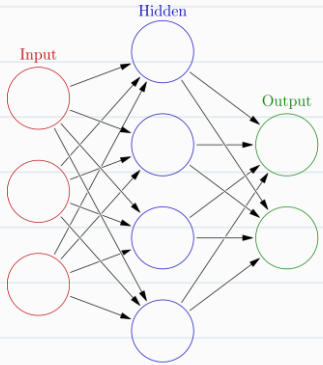
$$H(p, q_2) \approx 0.84$$

$$\boxed{H(p, q_1) < H(p, q_2)}$$

Conclusion: q_1 matches p better than q_2 ,hence it has lower cross entropy.

Why study Linear Algebra for ML?





Data **vector**: $\mathbf{x} \in \mathbb{R}^d$

- one data point with d features

Dataset **matrix**: $X \in \mathbb{R}^{n \times d}$

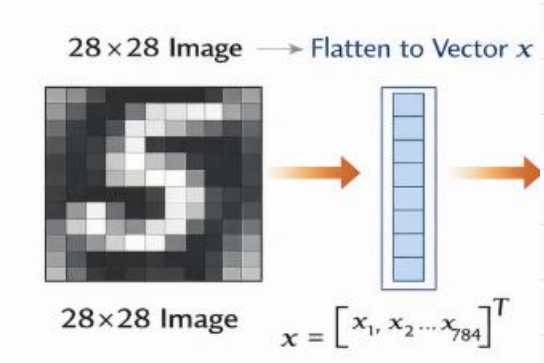
- rows = samples
- columns = features

Learning involves repeatedly applying transformations:

$$\mathbf{y} = \sigma(W\mathbf{x} + \mathbf{b}).$$

Matrix multiplication (Linear Transformtions)

Why study Linear Algebra for ML?

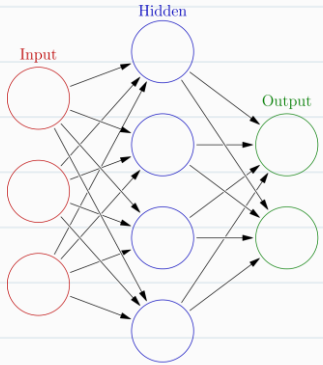


Data **vector**: $\mathbf{x} \in \mathbb{R}^d$

- one data point with d features

Dataset **matrix**: $X \in \mathbb{R}^{n \times d}$

- rows = samples
- columns = features



Matrix multiplication (Linear Transformtions)

Learning involves repeatedly applying transformations:

$\mathbf{y} = \sigma(W\mathbf{x} + \mathbf{b})$.

Lecture Content

- Length (Norm)
- Inner Product
- Linear Independence
- Orthogonality
- Span
- Basis
- Matrix multiplication as Linear Combination

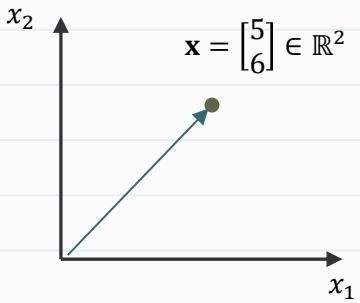
Vectors

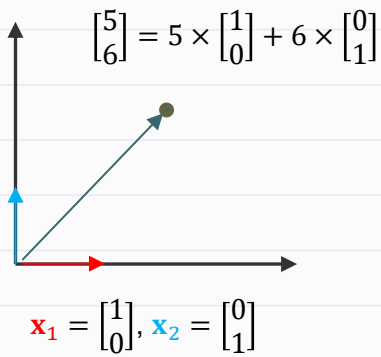
We work mostly with $V = \mathbb{R}^d$.

A vector:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} \in \mathbb{R}^d.$$

Example





Linear Combination

Given the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$, any vector

$$\mathbf{x} = \sum_{i=1}^{k} a_i \mathbf{x}_i, \quad a_i \in \mathbb{R}$$

is their linear combination.

Example

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 0 \\ 5 \end{bmatrix}$$

$$\mathbf{x} = 2 \times \mathbf{x}_1 + 3 \times \mathbf{x}_2$$

$$= 2 \times \begin{bmatrix} 1 \\ 2 \end{bmatrix} + 3 \times \begin{bmatrix} 0 \\ 5 \end{bmatrix}$$

$$= \begin{bmatrix} 2 \\ 19 \end{bmatrix}$$

(Informally) A vector space is a set of vectors where linear combinations of vectors stay within the set.

Span: Definition

For $S = \{\mathbf{x}_1, \dots, \mathbf{x}_k\}$, the span is defined as the set of all possible linear combinations

$$\text{span}(S) = \left\{ \sum_{i=1}^{k} a_i \mathbf{x}_i : \ a_i \in \mathbb{R} \right\}.$$

Geometric meaning:

- What is span $\left\{ \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right\}$? It is a line through origin with slope $\frac{1}{2}$.
- $\mathbf{x}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $\mathbf{x}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. Then $\text{span}\{\mathbf{x}_1, \mathbf{x}_2\}$ is the entire $\mathbb{R}^2$ (Why?)

Example

Let $\mathbf{x}_1 = (1,0,1)^T$,$\mathbf{x}_2 = (0,1,1)^T$.

Check if $\mathbf{z} = (1,1,2)^T$ lies in $\text{span}\{\mathbf{x}_1, \mathbf{x}_2\}$.

Are there $a, b \in \mathbb{R}$ that satisfy $a\mathbf{x}_1 + b\mathbf{x}_2 = \mathbf{z}$?

$$\text{a}\begin{bmatrix}1\\0\\1\end{bmatrix} + b\begin{bmatrix}0\\1\\1\end{bmatrix} = \begin{bmatrix}a\\b\\a+b\end{bmatrix} = \begin{bmatrix}1\\1\\2\end{bmatrix}$$

So $a = 1$, $b = 1$ and $a + b = 2$ matches. Hence $\mathbf{z} \in \text{span}\{\mathbf{x}_1, \mathbf{x}_2\}$.

Linear Independence

A set of vectors is linearly independent if **no vector in the set can be written as a linear combination of the others**.

Vectors $\mathbf{x}_1, \dots, \mathbf{x}_k$ are linearly independent iff

$$a_1\mathbf{x}_1 + \cdots + a_k\mathbf{x}_k = \mathbf{0} \Rightarrow a_1 = \cdots = a_k = 0.$$

Otherwise, dependent.

Interpretation: dependence = redundancy in representation.

Example:
In $\mathbb{R}^2$, let $\mathbf{x}_1 = (1,2)^\top, \mathbf{x}_2 = (2,4)^\top$. Are $\mathbf{x}_1, \mathbf{x}_2$ independent?

Example:

Are these three vectors in $\mathbb{R}^2$ independent?

$$\mathbf{v}_1 = \begin{pmatrix}1\\0\end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix}0\\1\end{pmatrix}, \quad \mathbf{v}_3 = \begin{pmatrix}1\\1\end{pmatrix}.$$

Clearly,

$\mathbf{v}_1$ and $\mathbf{v}_2$ independent.
$\mathbf{v}_1$ and $\mathbf{v}_3$ independent.
$\mathbf{v}_2$ and $\mathbf{v}_3$ independent.

So **every pair** among $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ is linearly independent.

However, the Three Together are Dependent because:

$$\mathbf{v}_3 = \mathbf{v}_1 + \mathbf{v}_2.$$

Equivalently,

$$1 \cdot \mathbf{v}_1 + 1 \cdot \mathbf{v}_2 + (-1) \cdot \mathbf{v}_3 = \mathbf{0},$$

and **not all coefficients are zero**.

Key takeaway:
Pairwise independence $\not\Rightarrow$ independence of the whole set.

What is the size of the largest independent set in $\mathbb{R}^2$?

Basis and Dimension

A set $\{\mathbf{b}_1, \ldots, \mathbf{b}_m\}$ is a basis of a vector space U if:

1. $U = \text{Span}\,(\{\mathbf{b}_1, \ldots, \mathbf{b}_m\})$
2. it is linearly independent

Then every $\mathbf{x} \in U$ has a *unique* representation

$$\mathbf{x} = \sum_{i=1}^{m} c_i \, \mathbf{b}_i.$$

The vector $[\mathbf{x}]_{\mathcal{B}} = (c_1, \ldots, c_m)^\top$ are the coordinates in basis $\mathcal{B}$.

Related facts:

Dimension of a vector space, $\dim(U)$ = number of vectors in any basis

Let U be any finite-dimensional vector space with $\dim(U) = m$, Then

- Any linearly independent set with exactly m vectors is a basis.
- Any set with fewer than m vectors is not a basis (it does not span U).
- No linearly independent set can have more than m vectors.

Why unique basis coordinates?

Given a basis, prove that the coordinates are unique for any vector.

Proof Outline: Assume there exist two representations:

$$\mathbf{x} = \sum_{i=1}^{m} c_i \, \boldsymbol{b}_i, \qquad \mathbf{x} = \sum_{i=1}^{m} a_i \, \boldsymbol{b}_i$$

Then equating them gives $\sum_{i=1}^{m}(c_i - a_i)\boldsymbol{b}_i = 0$.

Since, the vectors are linearly independent, this implies $c_i - a_i = 0, \ \forall i$.

In other words, $c_i = a_i, \forall i$ implying a unique representation.

Coordinates of a vector will depend on Basis

What will be the coordinates of the vector $\mathbf{x} = (2,2)^\top$ using the standard basis $\mathcal{E} = \{(1,0)^\top, (0,1)^\top\}$?

$$(2,2)^T = \mathbf{2} \times (1,0)^T + \mathbf{2} \times (0,1)^T$$

So $[\mathbf{x}]_{\mathcal{E}} = (\mathbf{2}, \mathbf{2})^\top$.

Changing the Basis:

Let $\mathbf{b}_1 = (1,1)^\top$ and $\mathbf{b}_2 = (1,-1)^\top$ ($\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2\}$ is another basis of $\mathbb{R}^2$)

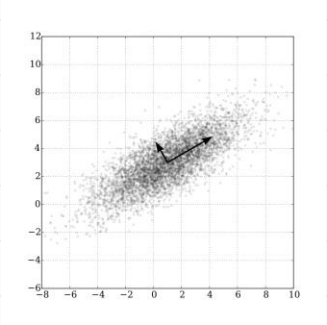
Represent $\mathbf{x} = (2,2)^\top$ as $\mathbf{x} = c_1\mathbf{b}_1 + c_2\mathbf{b}_2$:

$$c_1(1,1) + c_2(1,-1) = (c_1 + c_2,\ c_1 - c_2) = (2,2).$$

Solve:

$$c_1 + c_2 = 2, \quad c_1 - c_2 = 2 \Rightarrow \quad c_1 = 2,\ c_2 = 0.$$

So $[\mathbf{x}]_{\mathcal{B}} = (2,0)^\top$.



PCA

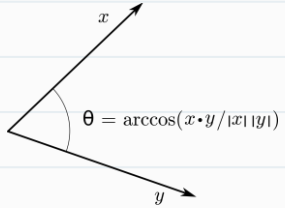
Remark:

- The coordinates of the vector will depend on the choice of basis
- Careful choice of basis leads to compact representations

Inner Product: Definition

In $\mathbb{R}^d$, **standard** inner product (dot product):

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y} = \sum_{i=1}^{d} x_i \, y_i = \|\mathbf{x}\| \|\mathbf{y}\| \, cos(\theta)$$



Length: $\|\mathbf{x}\| = \sqrt{\sum_i^d x_i^2}$

Angle: $\cos\theta = \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$

Other Inner products:

Weight each component differently:

$\langle \mathbf{x}, \mathbf{y} \rangle_W = \sum_i w_i \, x_i \, y_i$ (weights are positive)

Weight every pair differently:

$\langle \mathbf{x}, \mathbf{y} \rangle_A = \mathbf{x}^\mathbf{T} \mathbf{A} \, \mathbf{y} = \sum_{ij} A_{ij} x_i y_j$ ($\mathbf{A}$ is symmetric positive definite)

Example: Dot Product and Cosine Similarity

Let $\mathbf{x} = (1,2)^\top$,$\mathbf{y} = (2,1)^\top$.

$$\mathbf{x}^\top \mathbf{y} = 1 \cdot 2 + 2 \cdot 1 = 4.$$

$$\|\mathbf{x}\|_2 = \sqrt{1^2 + 2^2} = \sqrt{5}, \quad \|\mathbf{y}\|_2 = \sqrt{2^2 + 1^2} = \sqrt{5}.$$

Cosine Similarity: $\cos\theta = \dfrac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} = \dfrac{4}{5} \Rightarrow \theta \approx 36.9°.$

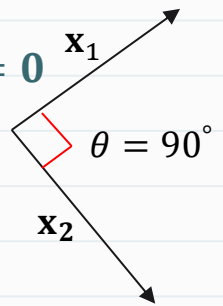
Interpretation:

high cosine similarity

⇒ vectors point in similar directions (common in word embeddings)

Orthogonal Vectors

$$< \mathbf{x}_1, \mathbf{x}_2 > = \mathbf{x}_1^\top \mathbf{x}_2 = 0$$



A set of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n\}$ are orthogonal if their pair-wise inner products is zero:

$$< \mathbf{x}_i, \mathbf{x}_j > = \mathbf{x}_i^\top \mathbf{x}_j = 0, \qquad \text{for i} \neq j$$

$$\cos\theta = \frac{\mathbf{x}_1^\top \mathbf{x}_2}{\|\mathbf{x}_1\| \|\mathbf{x}_2\|} = 0$$

$$\Rightarrow \theta = 90^\circ$$

Related Facts:

A zero vector is orthogonal to any vector since $\mathbf{x}^\top \mathbf{0} = \sum_i 0 \times x_i = 0$

Is there any relation between Orthogonality and Linear Independence?

For any set of non-zero vectors,

Orthogonal vectors ⇒ **Linearly Independent**
Linearly Independent ⇏ Orthogonal vectors

Prove this!

Examples

Example 1: $\mathbf{x} = (\mathbf{1}, \mathbf{0})^{\mathbf{T}}$, $\mathbf{y} = (\mathbf{0}, \mathbf{1})^{\mathbf{T}}$ **Orthogonal** and hence **linearly independent**

Example 2: $\mathbf{x} = (\mathbf{1}, \mathbf{1})^{\mathbf{T}}$, $\mathbf{y} = (\mathbf{1}, \mathbf{0})^{\mathbf{T}}$ **Linearly independent but NOT orthogonal**

Example 3: Let $\mathbf{x} = (1,2,1)^{\mathrm{T}}$ and $\mathbf{y} = (2, -1, 0)^{\mathrm{T}}$.

$$\mathbf{x}^{\mathrm{T}}\mathbf{y} = 1 \cdot 2 + 2 \cdot (-1) + 1 \cdot 0 = 2 - 2 + 0 = 0.$$

So $\mathbf{x} \perp \mathbf{y}$.

Orthonormal Basis

A set of vectors is **orthonormal** if it is orthogonal and each vector has unit norm.

Orthonormal Basis: An orthonormal basis is a basis whose vectors are orthonormal.

Important Fact:
In an orthonormal basis, **coordinates are just inner products**.

If $\{\mathbf{b}_1, \ldots, \mathbf{b}_d\}$ is an orthonormal basis for $\mathbb{R}^d$, then any vector $\mathbf{x} \in \mathbb{R}^d$ can be expressed as:

$$\mathbf{x} = \sum_i < \mathbf{x}, \mathbf{b_i} > \mathbf{b}_i$$

Example

Represent $\mathbf{x} = \begin{pmatrix} 3 \\ 1 \end{pmatrix} \in \mathbb{R}^2$ using the basis

$$\mathcal{B} = \left\{ \mathbf{v}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \mathbf{v}_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right\}.$$

Method 1: General Method (Solve Linear Equations) $\mathbf{x} = c_1\ \mathbf{v_1} + c_2\ \mathbf{v_2}$

$$\begin{pmatrix} 3 \\ 1 \end{pmatrix} = \frac{c_1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \frac{c_2}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Equating components gives

$$\frac{c_1 + c_2}{\sqrt{2}} = 3, \qquad \frac{c_1 - c_2}{\sqrt{2}} = 1.$$

Solving these equations gives the coordinates.

Example Contd.

Represent $\mathbf{x} = \begin{pmatrix} 3 \\ 1 \end{pmatrix} \in \mathbb{R}^2$ using the basis

$$\mathcal{B} = \left\{ \mathbf{v}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \mathbf{v}_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right\}.$$

Verify Orthonormality: Indeed, $\| v_1 \| = \| v_2 \| = 1$ and $v_1^{\mathsf{T}} v_2 = 0$.

Method 1: Using Inner product method $\mathbf{x} = c_1 \mathbf{v_1} + c_2 \mathbf{v_2}$

$$c_1 = \langle \mathbf{x}, \mathbf{v_1} \rangle = \frac{1}{\sqrt{2}}(3 + 1) = 2\sqrt{2},$$

$$c_2 = \langle \mathbf{x}, \mathbf{v_2} \rangle = \frac{1}{\sqrt{2}}(3 - 1) = \sqrt{2}.$$

Hence

$$[x]_{\mathcal{B}} = \begin{pmatrix} 2\sqrt{2} \\ \sqrt{2} \end{pmatrix}.$$

Norms: Definition and Examples

A **norm** $\|\cdot\|$ satisfies:

- Non– negative: $\|\mathbf{x}\| \geq 0,$
- Zero length only for the zero vector: $\|\mathbf{x}\| = 0 \Leftrightarrow \mathbf{x} = \mathbf{0},$
- Scaling property: $\|a\mathbf{x}\| = |a|\|\mathbf{x}\|,$
- Triangle inequality $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|.$

Examples:

$$\|\mathbf{x}\|_2 = \sqrt{\sum_i x_i^2}, \quad \|\mathbf{x}\|_1 = \sum_i |x_i|, \quad \|\mathbf{x}\|_\infty = \max_i |x_i|.$$

Let $\mathbf{x} = (3, -4, 12)^\mathsf{T}$.

$$\|\mathbf{x}\|_2 = \sqrt{3^2 + (-4)^2 + 12^2} = 13.$$
$$\|\mathbf{x}\|_1 = |3| + |-4| + |12| = 19,$$
$$\|\mathbf{x}\|_\infty = \max\{3,4,12\} = 12.$$

Intuition (Helpful to prevent Over fitting):

L2: encourages small energy

L1: encourages sparsity;

L∞: controls worst-case coordinate.

Matrix Multiplication

$$A = \begin{bmatrix} 2 & 4 \\ 2 & 3 \\ 3 & 7 \end{bmatrix}, \qquad \mathrm{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

Matrix–vector multiplication produces

$$A\mathrm{x} = \begin{bmatrix} 2x_1 + 4x_2 \\ 2x_1 + 3x_2 \\ 3x_1 + 7x_2 \end{bmatrix}.$$

Row view (component-wise):
Each entry of $A\mathrm{x}$ is a dot product of a row of A with x.

Key question:
Is there a more geometric way to understand this operation?

Matrix Multiplication

Column view: Write the columns of $A = \begin{bmatrix} 2 & 4 \\ 2 & 3 \\ 3 & 7 \end{bmatrix}$ as

$$\mathrm{v}_1 = \begin{bmatrix} 2 \\ 2 \\ 3 \end{bmatrix}, \qquad \mathrm{v}_2 = \begin{bmatrix} 4 \\ 3 \\ 7 \end{bmatrix}.$$

Then

$$\boxed{A\mathrm{x} = \begin{bmatrix} 2x_1 + 4x_2 \\ 2x_1 + 3x_2 \\ 3x_1 + 7x_2 \end{bmatrix} = x_1\mathbf{v}_1 + x_2\mathbf{v}_2}$$

$\mathbf{x}$ provides the coordinates
columns of A provide the directions
$A\mathbf{x}$ is a linear combination of columns

This is the a useful viewpoint in ML.

Thank You